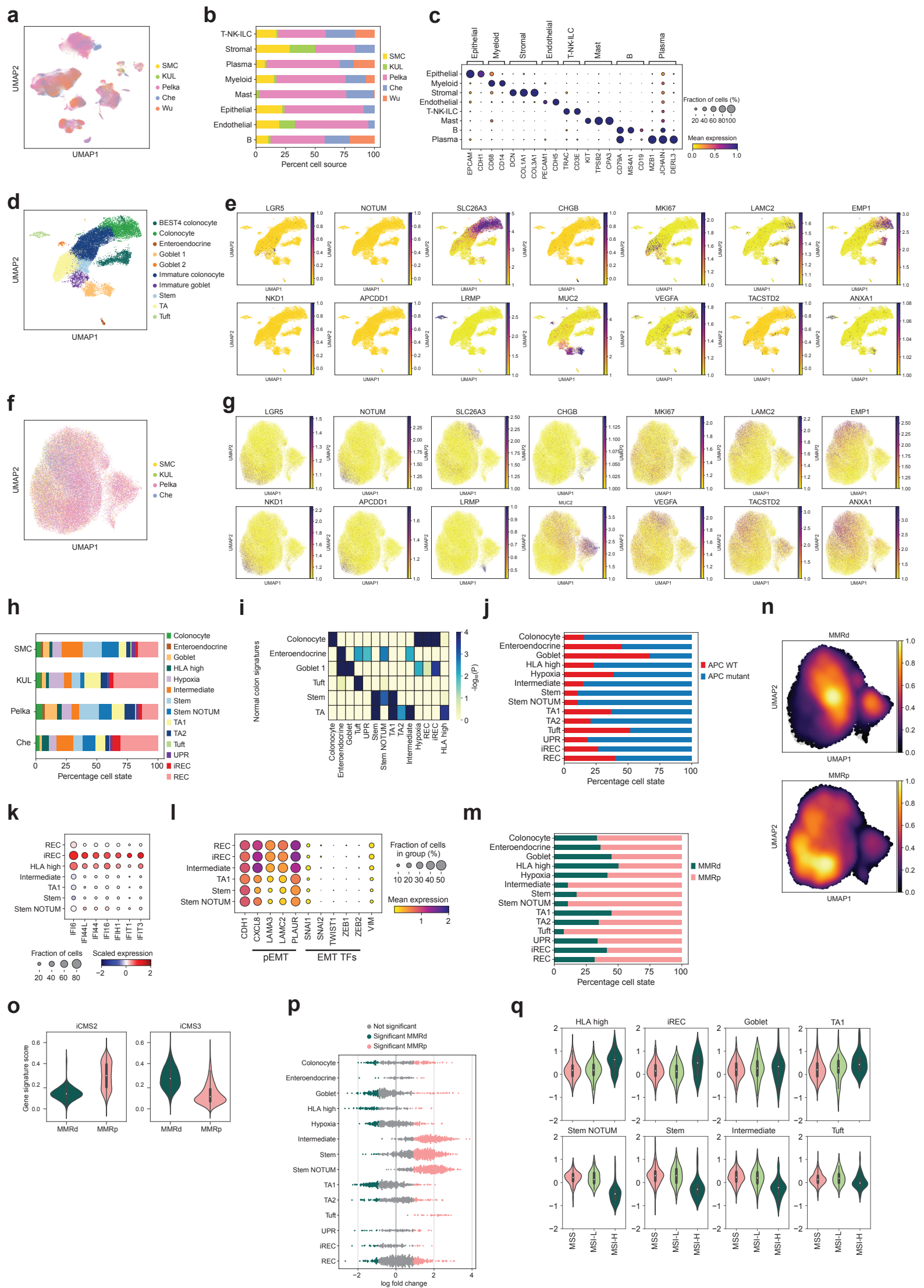
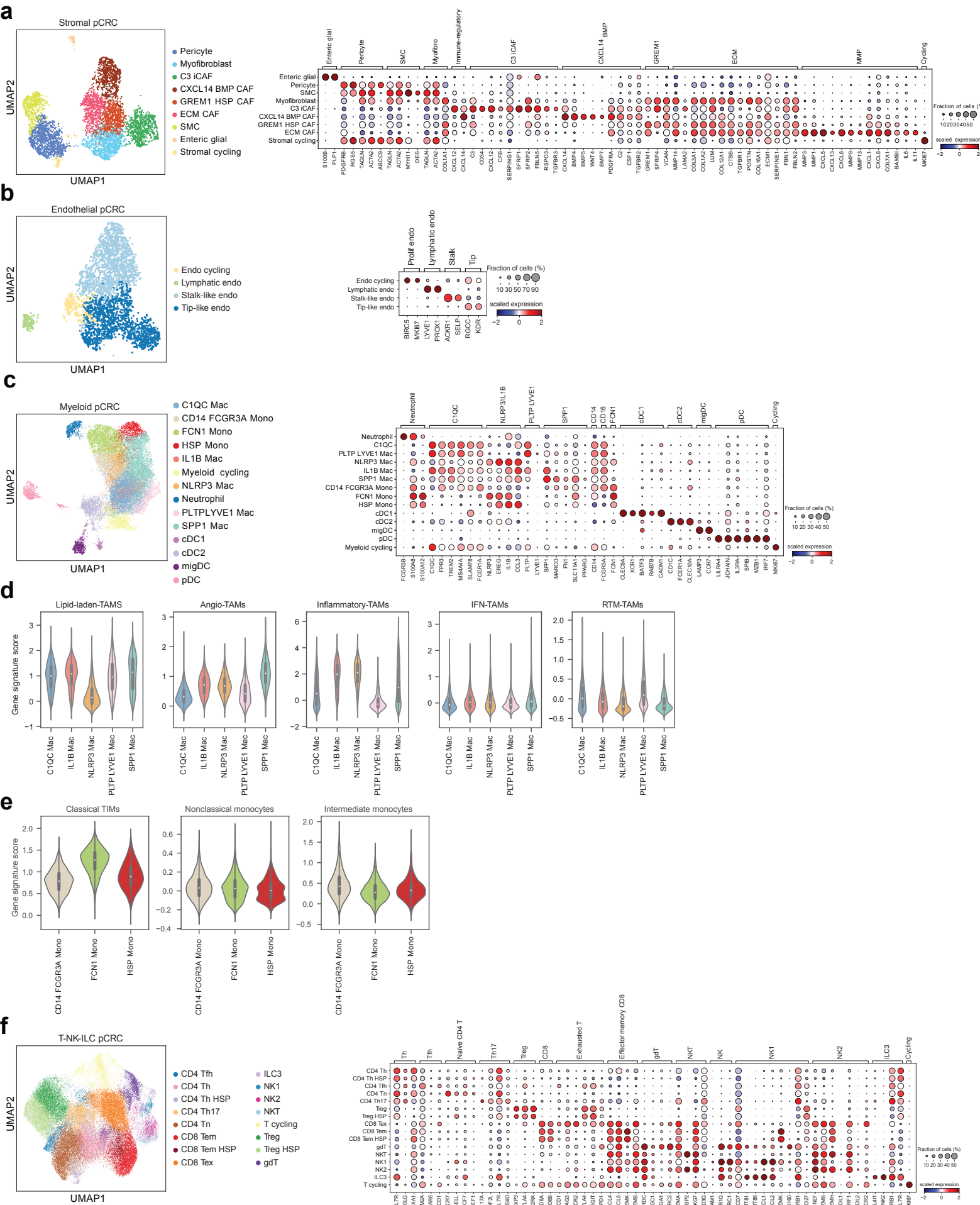


Extended Data Figures



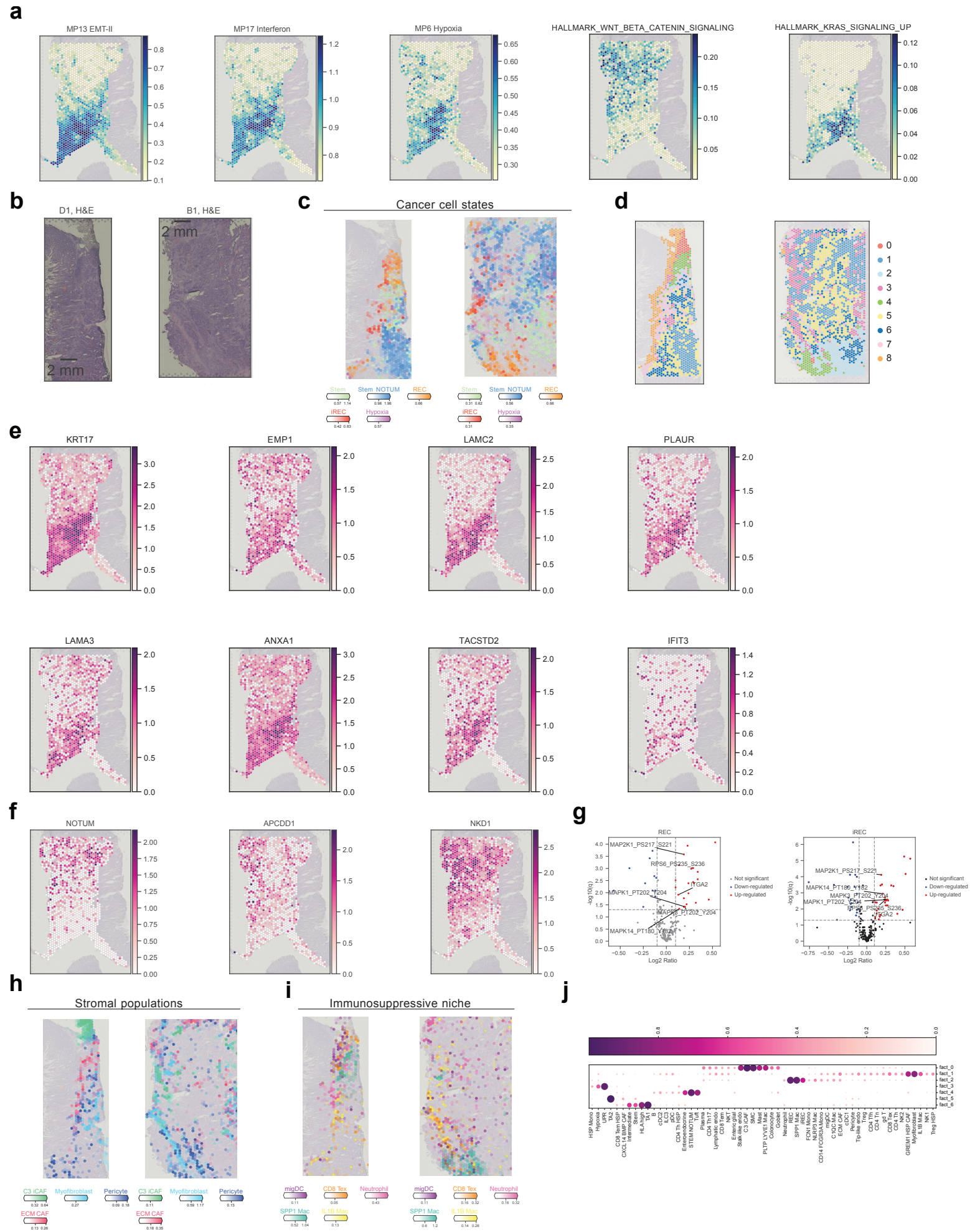
Extended Data Fig. 1: Cancer cell states in primary CRC (pCRC).

a. UMAP representation of integrated publicly available pCRC scRNA-seq data^{18–20,29}. **b.** Proportions of the major cell types in pCRC tumours across the five cell sources. **c.** Dotplot showing the average expression of discriminative marker genes of each major cell type (mean log₁₀ normalised counts). The colour scheme represents the mean expression from 0 to 1. The dot size indicates the fraction of cells expressing each marker in each cell type. **d.** UMAP representation of integrated publicly available normal colon epithelial cell scRNA-seq data^{19,95,96}. **e.** UMAP representation of normal colon epithelial cells showing the expression (mean log₁₀ normalised counts) of cancer cell state marker genes. **f.** UMAP representation of integrated malignant pCRC cells showing the dataset of origin. **g.** UMAP representation of malignant pCRC cells showing the expression (mean log₁₀ normalised counts) of the indicated genes. **h.** Proportions of cancer cell states present in pCRC datasets. **i.** GEA in pCRC cell states of the expression of normal colon cell type signatures. The top 100 DEGs ranked by log₂ fold change for the indicated normal colon cell type were used as signatures. **j.** Proportions of cancer cell states in *APC* WT or *APC* mutant tumours in SMC and KUL datasets. **k.** Dotplot showing the scaled expression of ISGs in each of the indicated cancer cell states. **l.** Dotplot showing the expression (mean log₁₀ normalised counts) of partial EMT (pEMT) genes, EMT transcription factors (TFs) and VIM in the indicated cancer cell states and stromal cells. **m.** Proportions of mismatch repair deficient (MMRd) and mismatch repair proficient (MMRp) cells in each pCRC cell state. **n.** Density of cells in UMAP representation of MMRd (up) and MMRd (down) pCRC cells. Density is scaled between 0 and 1. **o.** Gene signature scores for iCMS2 and iCMS3 signatures²⁸ in MMRd and MMRp pCRC cells. **p.** Beeswarm plot showing the log fold change in abundance of MMRd and MMRp cells in pCRC cell state neighbourhoods determined using Milo¹¹⁷. Each point is a cellular neighbourhood from a KNN-graph. Neighbourhoods with a spatial false discovery rate (FDR) < 0.05 were considered as statistically significant. **q.** Expression of gene signatures for the indicated cancer cell states in TCGA bulk RNA-seq samples whether tumours are microsatellite stable (MSS, n = 185), microsatellite instability low (MSI-L, n = 43) or microsatellite instability high (MSI-H, n = 36). Gene signatures were obtained from the top 50 DEGs ranked by log₂ fold change from analysis of pCRC malignant states and TME subpopulations (Supplementary Table 3).



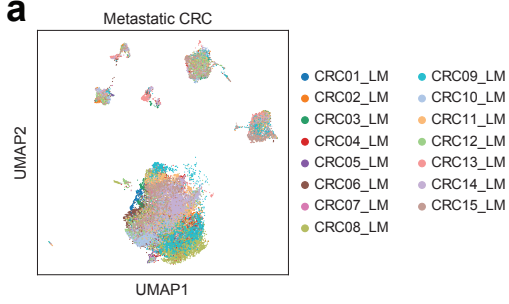
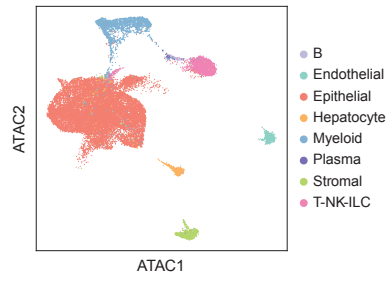
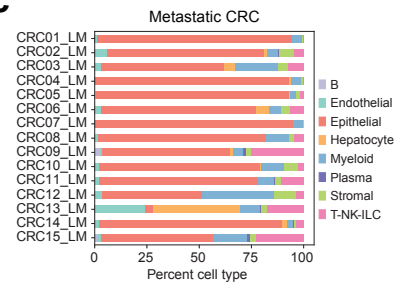
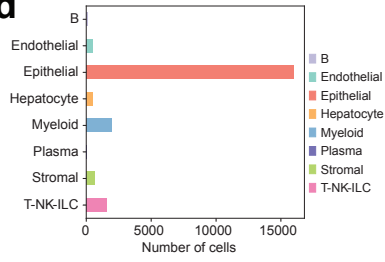
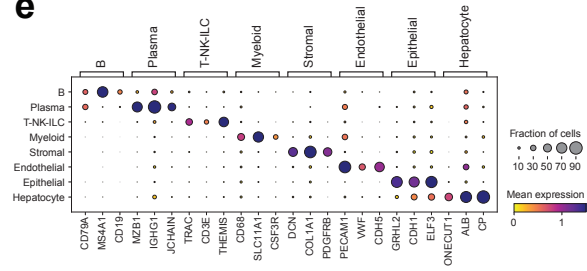
Extended Data Fig. 2. scRNA-seq analysis of the tumour microenvironment in primary CRC (pCRC) upon integration of publicly available datasets.

a. UMAP representation of stromal subpopulations (left). Dotplot showing the scaled expression of marker genes of fine-grained stromal subpopulations (right). **b.** UMAP representation of endothelial subpopulations (left). Dotplot showing the scaled expression of marker genes of fine-grained endothelial subpopulations (right). **c.** UMAP representation of myeloid subpopulations (left). Dotplot showing the scaled expression of marker genes of fine-grained myeloid subpopulations (right). **d.** Violin plots depicting gene signature scores derived from signatures of recurrent tumour-associated macrophage (TAM) subsets obtained from a single-cell RNA-seq analysis spanning over 15 tumour types (including CRC)³⁷ in the identified TAM subpopulations. In particular, we included signatures of lipid-laden TAMs (Lipid-laden-TAMs), pro-angiogenic TAMs (Angio-TAMs), inflammatory cytokine-enriched TAMs (Inflammatory TAMs), interferon-primed TAMs (IFN-TAMs), resident-tissue macrophages (RTM-TAMs). **e.** Violin plots depicting scores for gene signatures of tumour-infiltrating monocyte subsets obtained from a single-cell RNA-seq analysis spanning over 15 tumour types (including CRC)³⁷ in the identified monocyte subpopulations. **f.** UMAP representation of T-NK-ILC subpopulations (left). Dotplot showing the scaled expression of marker genes of fine-grained T-NK-ILC subpopulations (right). Tfh: follicular helper T cell, Th: helper T cell, Tn: naïve T cell, Tem: effector memory T cell, Tex: exhausted T cell, NK: natural killer cell, Treg: regulatory T cell, gdT: gamma delta T cell, ILC: innate lymphoid cell, Mac: macrophage, Mono: monocyte, DC: dendritic cell, CAF: cancer-associated fibroblast, SMC: smooth muscle cell, Endo: endothelial.



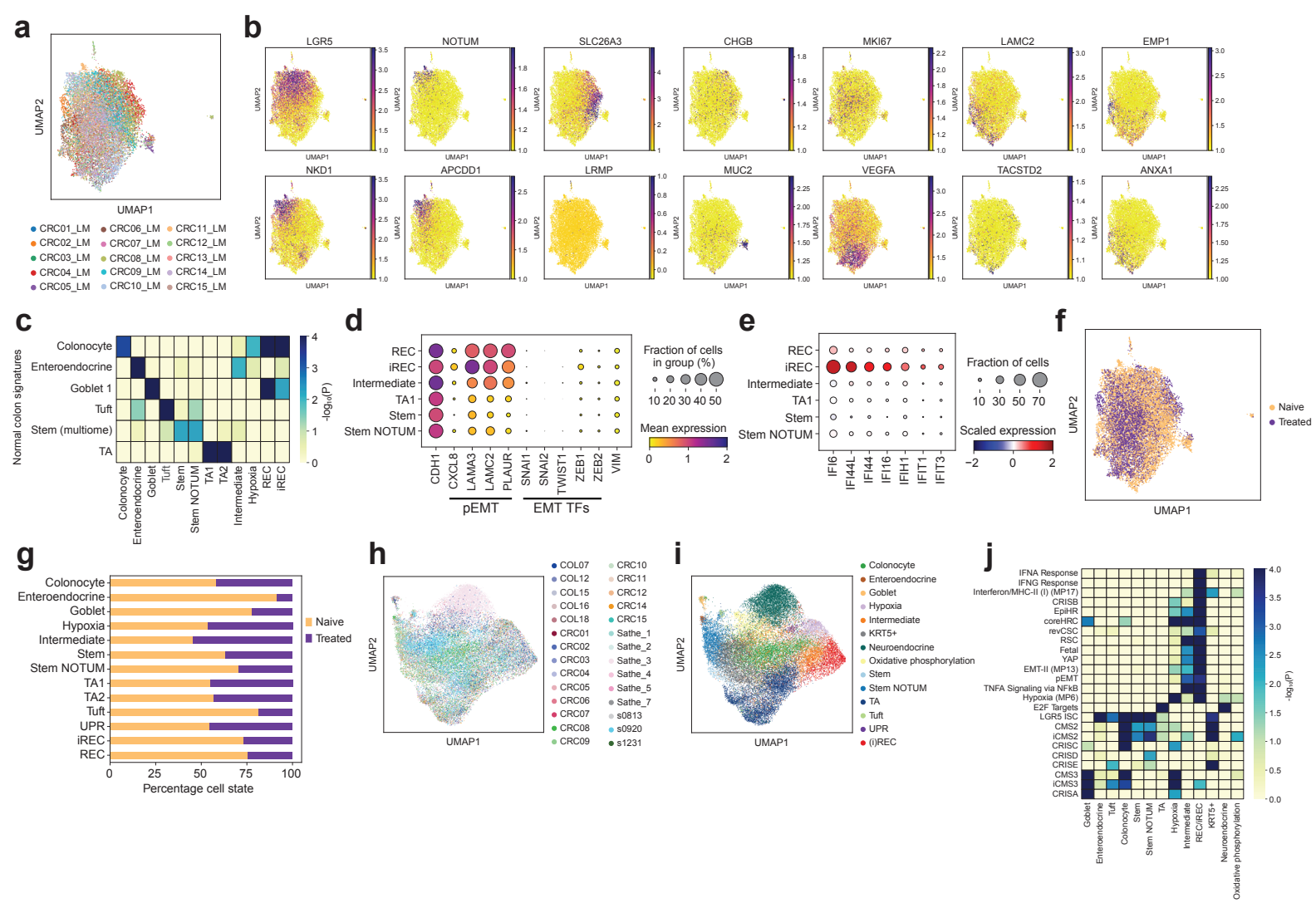
Extended Data Fig. 3: Spatial mapping of cancer cell states and identification of neighbourhoods in primary CRC using publicly available 10X Visium data.

a. Gene expression signature scores of spots for five different signatures. EMT, interferon response and hypoxia cancer cell signatures are derived from a single-cell RNA-seq analysis of 24 tumour types (including CRC)¹⁰⁰. Signatures for WNT signalling and KRAS signalling are obtained from MSigDB Hallmarks. **b.** H&E staining of primary colorectal cancer samples D1 and B1. **c.** Cell abundance estimates of cancer cell states across spatial locations of two primary CRC samples D1 and B1 capturing the tumour core and invasive edge. **d.** Spatial cellular neighbourhood identification, shown in samples D1 and B1. **e.** Gene expression (mean log1p normalised counts) of cancer state-specific marker genes associated with REC and iREC states in representative sample A1. **f.** Gene expression (mean log1p normalised counts) of Stem NOTUM-specific marker genes in representative sample A1. **g.** Reverse Phase Protein Array (RPPA) data showing differential abundance of phospho-peptides/proteins in TCGA bulk CRC tumours. Tumours with high expression of iREC (left) or REC (right) signatures were compared to tumours with low expression of the respective gene signature. A positive log2 ratio indicates the phospho-peptide/protein had higher abundance in tumours with higher expression of the signatures. Statistical significance was determined using a Student's T test with Benjamini-Hochberg correction. **h.** Abundance of stromal cell states in D1 and B1. **i.** Abundance of immune cell states in the cellular neighbourhood containing iREC and REC cancer cell states in D1 and B1. **j.** Cellular neighbourhood identification using the NMF module from cell2location. Dotplot depicts the relative factor loadings of each cell state, normalised across factors for each cell state. Factors correspond to cellular neighbourhoods. Factor loadings are depicted by colour and dot size.

a**b****c****d****e**

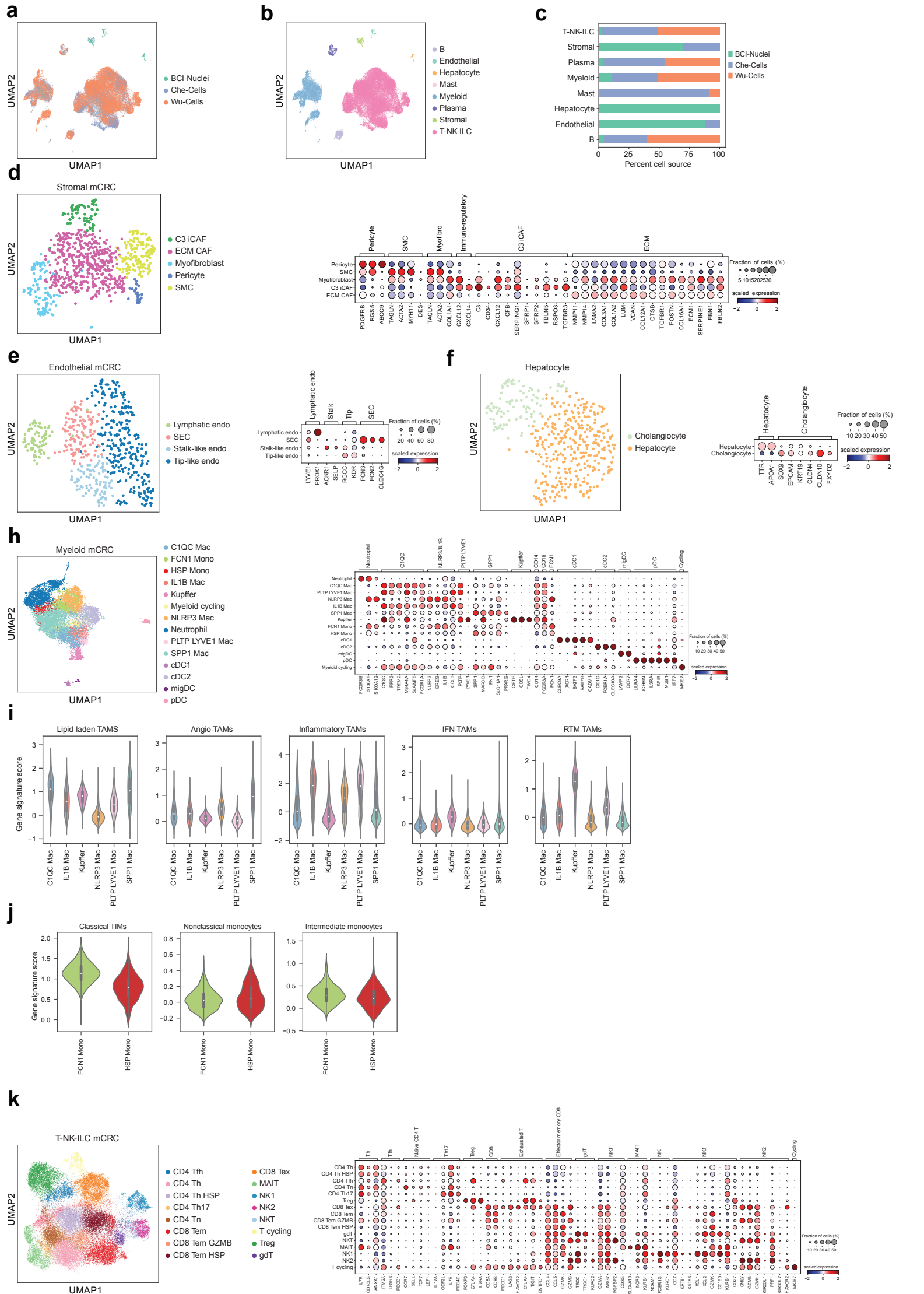
Extended Data Fig. 4: Integrated analysis of the single cell landscape of liver metastatic CRC.

a. UMAP representation of cells in the mCRC Multiome data, showing the sample of origin. UMAP is based upon RNA modality. **b.** UMAP representation of cell types in mCRC Multiome data. The representation is based upon the ATAC modality. **c.** Proportions of cell types in the mCRC Multiome dataset across patient samples. **d.** The number of cells of each cell type in mCRC Multiome data. **e.** Dotplot showing the expression of marker genes in metastatic CRC. Mean log1p normalised counts for each cell type is shown.



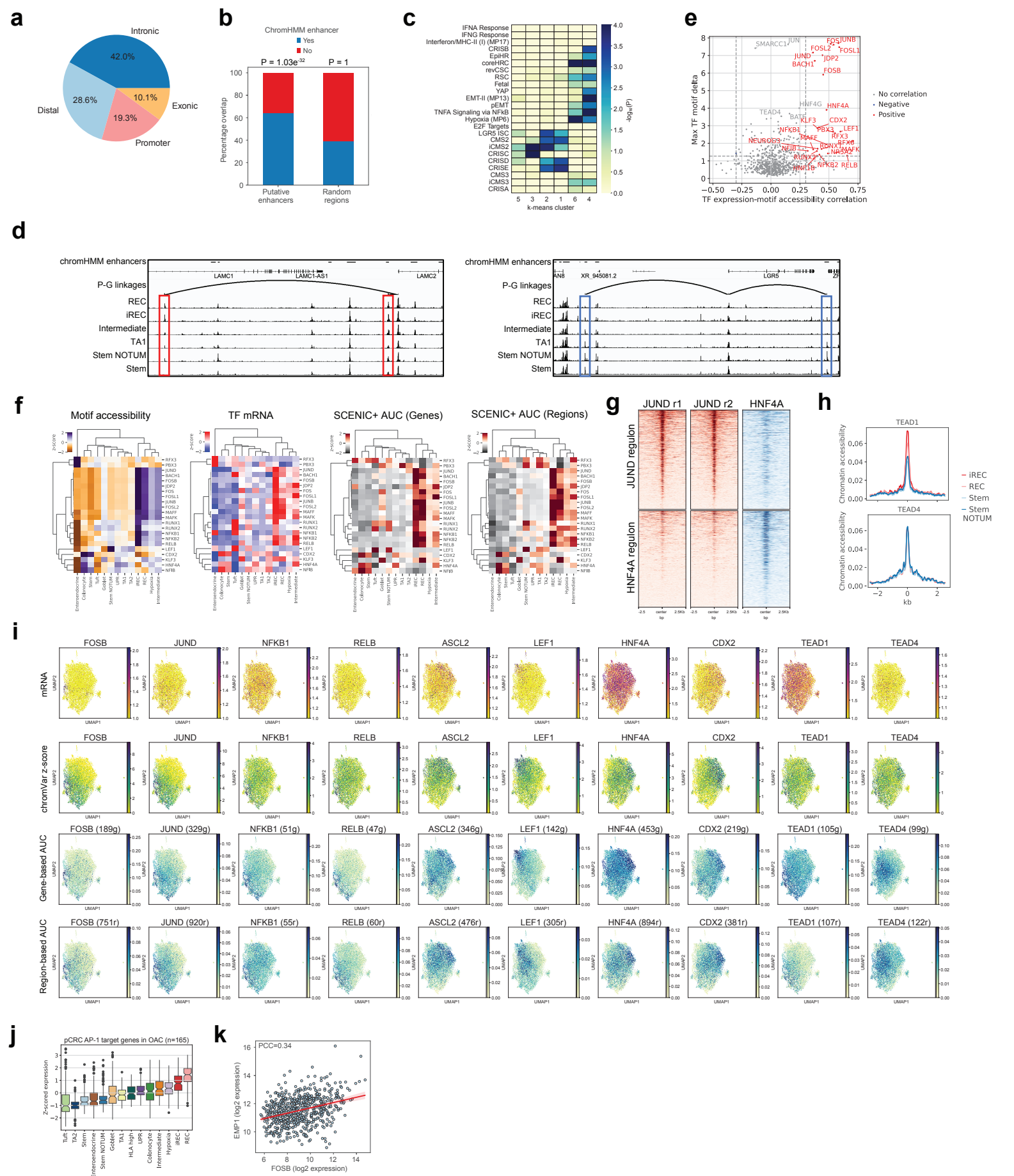
Extended Data Fig. 5: Cancer cell states in liver metastatic CRC.

a. UMAP representation of cancer cells in the mCRC Multiome data, showing the sample of origin. UMAP is based upon RNA modality. **b.** UMAP representation of cancer cells in mCRC, showing the expression (log1p normalised counts) of the indicated marker genes. **c.** GEA of normal colon cell type signature in mCRC cell states. **d.** Dotplot showing the expression (mean log1p normalised counts) of partial EMT (pEMT) genes, EMT transcription factors (TFs) and *VIM* in the indicated mCRC cell states and stromal cells. **e.** Dotplot showing the scaled expression of ISGs in each of the indicated mCRC cell states. **f.** UMAP representation of mCRC cells showing whether patients had received chemotherapy prior to specimen collection. UMAP is based upon RNA modality. **g.** Proportions of chemotherapy naïve or treated cells in mCRC cell states. **h.** UMAP representation showing the patient of origin of cancer cells in an analysis of our Multiome snRNA-seq data integrated with cancer cells in published mCRC scRNA-seq datasets REFS. **i.** Annotation of cancer cells in the integrated mCRC scRNA-seq analysis. **j.** Heatmap showing GEA of differentially expressed genes (DEGs) in the integrated scRNA-seq analysis. Genes in each signature are listed in Supplementary Table 2.



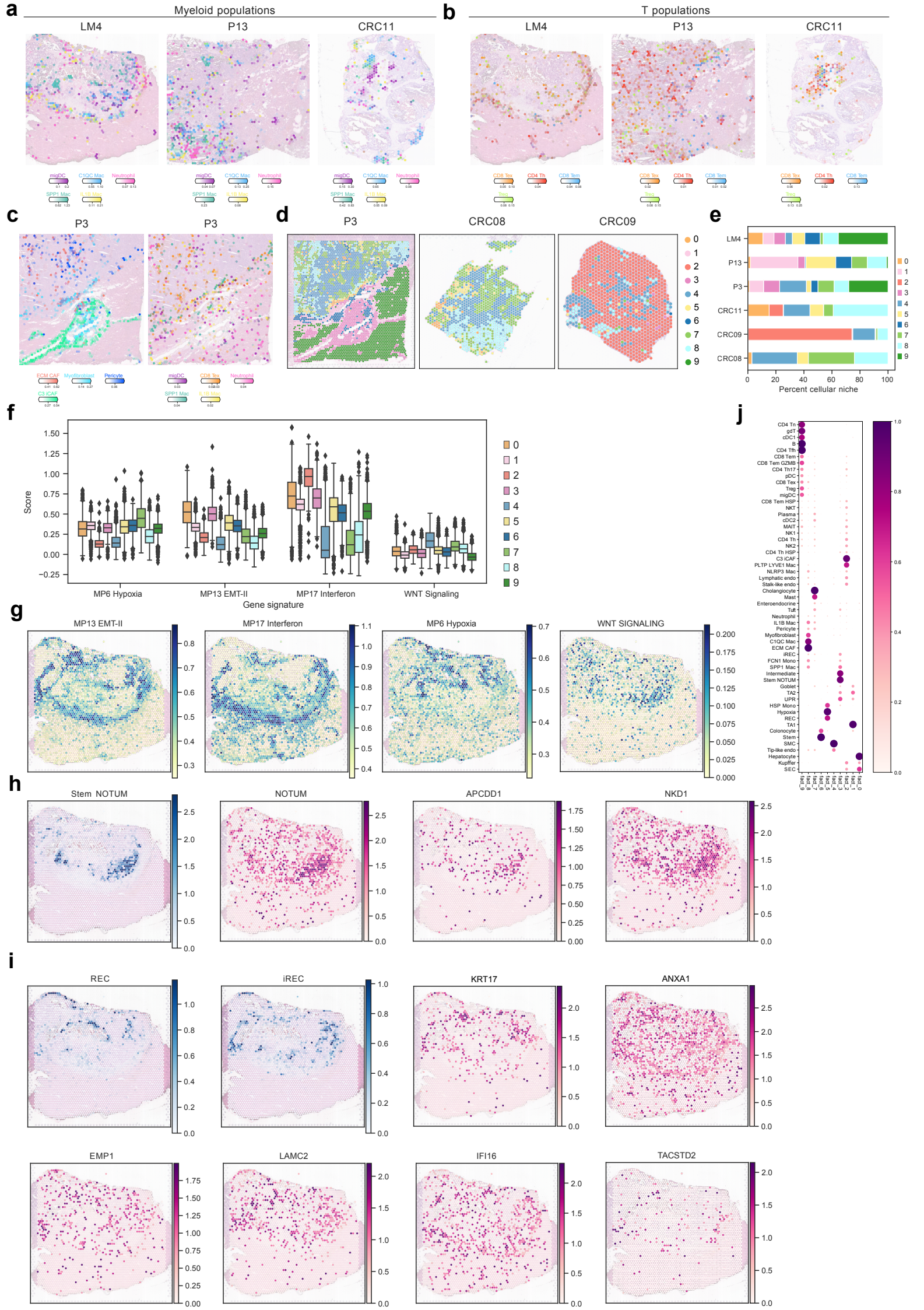
Extended Data Fig. 6: Characterisation of the tumour microenvironment in metastatic CRC (mCRC) upon integration of multiome and publicly available scRNA-seq datasets.

a. UMAP representation of integrated TME landscape from multiome and publicly available scRNA-seq data^{20,29} of liver mCRC tumours. **b.** UMAP representation of major cell types in the TME. **c.** Proportions of the major cell types in pCRC tumours across the three cell sources. **d.** UMAP representation of stromal subpopulations (right). Dotplot showing the scaled expression of marker genes of fine-grained stromal subpopulations (left). **e.** UMAP representation of endothelial subpopulations (left). Dotplot showing the scaled expression of marker genes of fine-grained endothelial subpopulations (right). **f.** UMAP representation of hepatocytes and cholangiocytes (left). Dotplot showing the scaled expression of marker genes of hepatocytes and cholangiocytes (right). **g.** UMAP representation of myeloid subpopulations (left). Dotplot showing the scaled expression of marker genes of fine-grained myeloid subpopulations (right). **h.** UMAP representation of myeloid subpopulations (left). Dotplot showing the scaled expression of marker genes of fine-grained myeloid subpopulations (right). **i.** Violin plots depicting gene signature scores derived from signatures of recurrent tumour-associated macrophages (TAM) subsets obtained from a single-cell RNA-seq analysis spanning over 15 tumour types (including CRC) in the identified TAM subpopulations³⁷. In particular, we included signatures of lipid-laden TAMs (Lipid-laden-TAMs), pro-angiogenic TAMs (Angio-TAMs), inflammatory cytokine-enriched TAMs (Inflammatory TAMs), interferon-primed TAMs (IFN-TAMs), resident-tissue macrophages (RTM-TAMs). **j.** Violin plots depicting scores for gene signatures of tumour-infiltrating monocyte subsets obtained from a single-cell RNA-seq analysis spanning over 15 tumour types (including CRC)³⁷ in the identified monocyte subpopulations. **k.** UMAP representation of T-NK-ILC subpopulations (left). Dotplot showing the scaled expression of marker genes of fine-grained T-NK-ILC subpopulations (right). SEC: sinusoidal endothelial cells.



Extended Data Fig. 7: Transcription factors regulating cancer cell states.

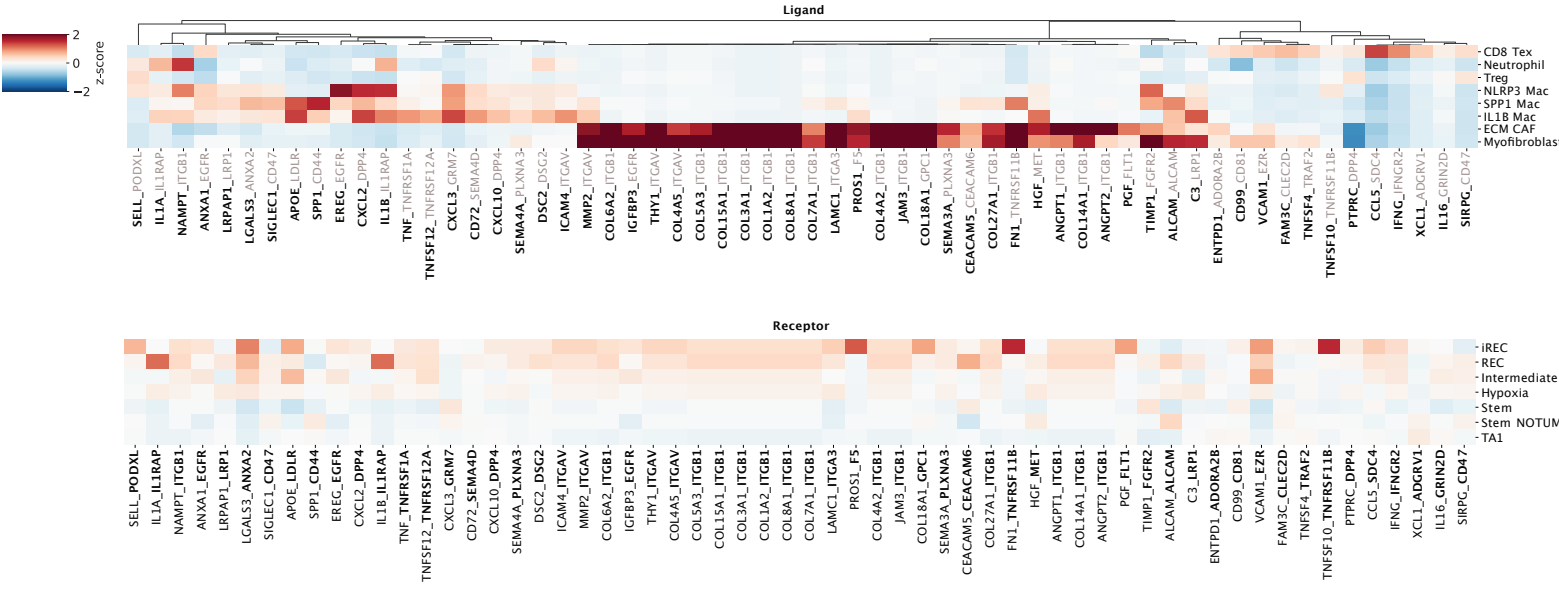
a. Genomic distribution of open chromatin regions ($n = 82,491$) in the CRC cell state union peakset. **b.** Percentage overlap of chromatin regions in P-G linkages ($n = 1,444$) or a random set of chromatin regions ($n = 1444$) with chromHMM enhancers⁴⁷ ($n = 33130$). **c.** GEA of genes in k means clusters for the indicated gene signatures. See Supplementary Table 2 for the gene signatures. **d.** Chromatin accessibility at *LAMC2* and *LGR5* loci in the indicated mCRC cell states. pEGLs are shown and chromHMM enhancers⁴⁷. **e.** Plot showing the maximum difference in chromVAR motif deviation z-score between CRC cell states, against correlation of chromVAR motif deviation and corresponding TF expression. **f.** Heatmaps showing (from left-right) z-scored mRNA expression, chromVAR motif deviation z-scores, enhancer driven regulons (eRegulons) z-scored SCENIC+ area under curve (AUC) scores for genes in regulons, eRegulons z-scored SCENIC+ AUC scores for chromatin regions in regulons of TFs highlighted in red in Extended Data fig. 7d. **g.** JUND and HNF4A ChIP-seq signal at chromatin regions in JUND¹⁰⁵ or HNF4A¹⁰⁶ eRegulons. r1 - replicate 1, r2 - replicate 2. **h.** Accessibility of chromatin regions in the indicated SCENIC+ regulons. **i.** UMAP representation showing TF mRNA expression, and corresponding chromVAR deviation z-scores, gene-based eRegulon AUC scores, region-based eRegulon AUC scores. **j.** mRNA expression levels of AP-1 target genes⁶¹ in pCRC cell states. **k.** Scatter plot showing the expression levels of *FOSB* and *EMP1* in bulk pCRC RNA-seq (TCGA) data ($n = 609$). PCC – Pearson correlation coefficient.



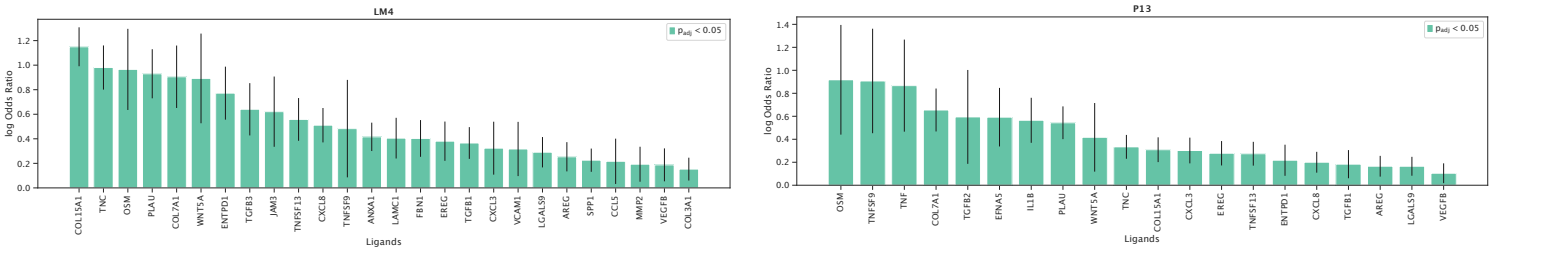
Extended Data Fig. 8: Spatial mapping of cancer cell states and identification of spatial cellular niches across liver metastatic CRC samples.

a. Estimated cell type abundances for distinct myeloid subsets across spatial locations of samples LM4, P13 and CRC11. **b.** Estimated cell type abundances for distinct T cell subsets across spatial locations of samples LM4, P13 and CRC11. **c.** Estimated cell type abundances for distinct stromal subpopulations (left) and immune subpopulations (right) in sample P3. **d.** Spatial cellular neighbourhoods in samples P3, CRC08 and CRC09. **e.** Proportions of the spatial cellular neighbourhoods across the six Visium samples. **f.** Boxplots depicting four gene expression signature scores in spatial cellular neighbourhoods. The boxplots show three quartiles (lower, median, upper) of the data distribution, with whiskers extending to data points that lie within 1.5 inter-quartile ranges of the lower and upper quartile. **g.** Gene expression signature scores of spots in representative sample LM4 for four different gene signatures¹⁰⁰: epithelial-to-mesenchymal transition (MP13 EMT-II), interferon response (MP17 Interferon), hypoxia (MP6 Hypoxia) and Hallmark WNT Beta Catenin Signalling (WNT Signaling). **h.** Gene expression (mean log1p normalised counts) of Stem NOTUM-specific marker genes in representative sample LM4. **i.** Gene expression (mean log1p normalised counts) of cancer state-specific marker genes associated with REC and iREC states in representative sample LM4. **j.** Cellular neighbourhood identification using the NMF module from cell2location. Dotplot depicts the relative factor loadings of each cell state, normalised across factors for each cell state. Factors correspond to cellular neighbourhoods. Factor loadings are depicted by colour and dot size.

a



b



Extended Data Fig. 9: Spatially resolved cell-cell interactions in the cellular neighbourhood surrounding interferon response REC cancer cell state.

a. Heatmaps summarising the inferred spatial cell-cell interactions mediated by stromal, myeloid, CD8 exhausted T (CD8 Tex) and regulatory T cells (Treg) in the cellular neighbourhood containing iRECs, with iRECs as the receiver, using CellPhoneDB and NicheNet. Specifically, we identified potential upstream ligand-receptor pairs which can induce the NF- κ B (RELB) regulon program in the neighbouring pro-metastatic phenotype. Z-score of the gene expression of selected potential ligands in each cell type of the cellular neighbourhood (top panel) and z-score of gene expression of corresponding receptors in cancer cell states (bottom panel). In both heatmaps, the x-axis denotes ligand-receptor interactions, with the ligand in bold and receptor in grey for sender cells (top panel) and the ligand in grey and the receptor in bold for cancer cell states (bottom panel). **b.** Ligand spatial enrichment in the cellular neighbourhood surrounding iREC cancer state in two representative samples (LM4 and P13). Ligands predicted to induce AP-1 regulon activation and statistically significant enrichments are shown. Barplot showing statistically significant enrichments, i.e. positive log odds ratio associated with adjusted p-value < 0.05. Statistical significance is assessed with Chi-square test using a significance level of 5%. The p-values are adjusted for multiple testing using the Benjamin-Hochberg correction method. The 95% confidence interval for the log odds ratio is plotted.